From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference

Year: 2023 | *Citations: 233* | *Authors: S. Samsi, Dan Zhao, Joseph McDonald*

Abstract

Large language models (LLMs) have exploded in popularity due to their new generative capabilities that go far beyond prior state-of-the-art. These technologies are increasingly being leveraged in various domains such as law, finance, and medicine. However, these models carry significant computational challenges, especially the compute and energy costs required for inference. Inference energy costs already receive less attention than the energy costs of training LLMs-despite how often these large models are called on to conduct inference in reality (e.g., ChatGPT). As these state-of-the-art LLMs see increasing usage and deployment in various domains, a better understanding of their resource utilization is crucial for cost-savings, scaling performance, efficient hardware usage, and optimal inference strategies. In this paper, we describe experiments conducted to study the computational and energy utilization of inference with LLMs. We benchmark and conduct a preliminary analysis of the inference performance and inference energy costs of different sizes of LLaMA-a recent state-of-the-art LLM-developed by Meta AI on two generations of popular GPUs (NVIDIA V100 & A100) and two datasets (Alpaca and GSM8K) to reflect the diverse set of tasks/benchmarks for LLMs in research and practice. We present the results of multi-node, multi-GPU inference using model sharding across up to 32 GPUs. To our knowledge, our work is the one of the first to study LLM inference performance from the perspective of computational and energy resources at this scale.